IES Institute of Education Sciences

APPENDIXES
October 2024

Regional Educational Laboratory Mid-Atlantic
At Mathematica

Stabilizing School Performance Indicators in New Jersey to Reduce the Effect of Random Error

See https://ies.ed.gov/ncee/rel/Products/Publication/108130 for the full report.

Appendix A. Theoretical review

The Regional Educational Laboratory (REL) Mid-Atlantic study team used a stabilization model called Bayesian hierarchical modeling to stabilize school accountability data for the New Jersey Department of Education (NJDOE). NJDOE provided these data at the student subgroup[1] level for each school. The model was designed to reduce measurement error and improve the reliability of performance indicators, especially by reducing measurement error in scores from small student groups. Stabilization considers data for each student group in the broader context of data from all student groups within the same subgroup and data from the same student group in other years and then applies structured assumptions about how these scores relate to each other to reduce the effects of measurement error on performance indicators. This increases the reliability of performance indicators, especially for small student groups, where measurement error can have an outsized effect.

A major challenge in measuring academic performance is that all indicator scores are subject to measurement error, which can undermine the reliability of results. For smaller schools and student groups, error can cause performance indicators to fluctuate excessively from year to year and occasionally cause scores to fall below a cutoff for designating a school or student group for Comprehensive Support and Improvement (CSI) or Additional Targeted Support and Improvement (ATSI) due solely to measurement error. To mitigate this, states choose a minimum size for student groups to be considered in accountability processes. However, this introduces a tradeoff between accuracy and equity, because larger minimum group sizes provide more reliable scores but make small student groups "invisible," so that they might not receive support if they need it. To explore stabilization as a means to circumvent the tradeoff and improve score reliability across all student group

[1] The Every Student Succeeds Act of 2015 defines student groups within a school using demographic designations, including race/ethnicity, students with disabilities, economically disadvantaged students, and English learner students. This report refers to these designations as *subgroups* and refers to students of a particular designation within each school as *student groups*, sometimes shortened to *groups*. This distinction is useful when discussing scores and group size distributions within designations.

sizes, the study team applied a stabilization model to all NJDOE performance indicators for all student groups (group sizes as small as 10 students) with publicly available performance data.

To illustrate the usefulness of borrowing information across entities, consider a hypothetical school with a relatively small population of students of two or more races, where group size ranges between 15 and 25 students, depending on the year. Suppose that these students typically perform similarly to their peers, so that their measured performance over time closely tracks that of the whole school. This is not surprising, as these students share many important experiences and resources with their peers, including teachers, administrators, curricula, and facilities. Suppose, however, that in one year several students in this group do not test well for any number of reasons unrelated to their true content mastery, including test anxiety, having a poor night's sleep, or fighting with their parents before school. In that year, the average test score for the subgroup is much lower than the average for the whole school, but not because the students' content mastery is truly worse. Stabilizing the score for students of two or more races in that year by borrowing information from the average performance in School X and the same group's performance in previous years reduces the effect of the error-affected scores on the group average. A group's performance may be stabilized in a variety of ways, including borrowing information from the same subgroup in other schools in the same year or across time.

The methods used in this study are grounded in evidence from the statistical literature, which establishes that stabilization reduces measurement error and may do so better than using a simple average (Efron & Morris, 1977). This finding is a mainstay of statistical analysis. Stein (1956) showed, mathematically, that stabilizing performance estimates by borrowing information across entities (such as schools, hospitals, or student groups) yields more accurate estimates than a simple average. This idea was formalized in the James-Stein estimator (James & Stein, 1961) and has been expanded upon by many subsequent studies. This method of borrowing information across entities allows similar entities to provide contextual evidence for one another in the absence of larger bodies of information from the specific entity. Doing so reduces variance in estimates that is partially caused by noise from measurement error, effectively increasing the signal-to-noise ratio of an estimate and allowing estimates to better represent true values with less information.

Stein's work on stabilization has been applied to performance measurement applications across fields, including education (teacher evaluation) and health care (hospital quality measurement programs) (Mulhern & Opper, 2022). A report to the Centers for Medicare and Medicaid Services on hospital performance measurement endorsed the use of Bayesian stabilization for hospital-specific indicators, such as readmission and mortality rates (Ash et al., 2011). The authors observed that stabilization reduces the effect on data interpretation of random error on measurements and of regression to the mean—when noise in sampling produces one extreme value, the next measurement of that value is likely to be less extreme. Regression to the mean can lead to interpretations of a sequence of measurements as meaningful when they are not. By reducing the effects of measurement error, stabilization produces more reliable data that decisionmakers can interpret with greater confidence and clarity.

One feature of stabilization that may cause concern is that, by design, stabilization has a greater effect on estimates that are less precise, such as those arising from smaller student groups. When two estimates of the same unstabilized value are stabilized, one taken from a larger group and one from a smaller group, the stabilized estimate for the larger group will, on average, be closer to the unstabilized value than the stabilized estimate for the smaller group, which will be less precise.[2] Ash et al. (2011), in addressing concerns that this effect may mask the performance of smaller student groups, found minimal negative effects. Similarly, when Herrmann et al. (2016) investigated whether this effect would disproportionately reduce the rate at which teachers of small

[2] This is not guaranteed but reflects the fact that unstabilized estimates from larger groups are less susceptible to noise and will, therefore, be closer on average to the mean, given the standard deviation of the distribution.

groups of students were assigned consequences for poor performance, they found no statistically significant effect.[3]

Taken together, these studies indicate that stabilization has great potential to improve the reliability of performance indicators by reducing the effect of measurement error without inhibiting correct identification of schools in need of assistance. This study examined how the stabilization of indicators used in school accountability (including proficiency rates and median student growth percentiles) affects performance estimates and school improvement classifications in New Jersey. Similar to the study for the Pennsylvania Department of Education (Forrow et al., 2023), this study found favorable effects of Bayesian stabilization on reliability. The results of this study expand on those for Pennsylvania and are more generalizable to similar designations in other states, as New Jersey's process of using a summative score across performance indicators to determine CSI and ATSI classifications is more common than Pennsylvania's approach, which uses a multistage process for designating schools in need of support.

References

Ash, A. S., Fienberg, E., Louis, A., Norm, S. L. T., Stukel, A., & Utts, P. J. (2011). *Statistical issues in assessing hospital performance commissioned by the Committee of Presidents of Statistical Societies*. The COPSS-CMS White Paper Committee.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119-127.

Forrow, L., Starling, J., & Gill, B. (2023). *Stabilizing subgroup proficiency results to improve the identification of low-performing schools* (REL 2023-001). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/rel/Products/Publication/106926

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy, 3*(1), 1-10. https://doi.org/10.1080/2330443X.2016.1182878

James, W., & Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory* (pp. 443-460). New York, NY: Springer New York.

Mulhern, C., & Opper, I. M. (2022). Measuring and summarizing the multiple dimensions of teacher effectiveness (EdWorkingPaper 21-451). Annenberg Institute at Brown University. https://doi.org/10.26300/h9qh-0078

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-55, Vol. I*, 197-206. University of California Press, Berkeley.

[3] Note that the scope of this study was smaller, applying to a single school district rather than a state, and the results could differ if applied at a larger scale.

Appendix B. Methods

Data

The New Jersey Department of Education (NJDOE) provided cleaned and aggregated publicly available data at the school, year, performance indicator, and subgroup level for this study These data are also available via NJDOE's accountability website (New Jersey Department of Education, 2024). Different indicators apply to different subgroups, grades, and configurations of schools. Additionally, not all indicators were available in every year. A summary of which indicators appeared in which year and the school grades to which they applied are in table B1.

Table B1. Summary of data provided by the New Jersey Department of Education

Performance indicator	2015/16	2016/17	2017/18	2018/19	2021/22	Applicable to
Four-year high school graduation rate	X	X	X	X	X	High schools
Five-year high school graduation rate	X	X	X	X	X	High schools
English language arts proficiency	X	X	X	X	X	Grades 3-10
Math proficiency	X	X	X	X	X	Grades 3-10
English language arts growth		X	X	X		Grades 4-8
Math growth		X	X	X		Grades 4-7
English language proficiency progress			X	X	X	All grades[a]
Chronic absenteeism		X	X	X	X	All grades

Note: Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

a. Applies to English learner students only.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

The study team analyzed each indicator schoolwide (except for English language proficiency progress) and for each of the following subgroups: racial/ethnic categories (Asian, Native Hawaiian, or Pacific Islander students, Black or African American students, Hispanic students, White students, Native American students, and multiracial students); economically disadvantaged students; students with disabilities; and students who were identified as English learners. NJDOE also provided data on grade span and Title I designations for each school in each year. The study team used grade span data to identify the indicators expected for each school as well as data on Title I status for the Comprehensive Support and Improvement (CSI) and Additional Targeted Support and Improvement (ATSI) analysis. The key variables provided included student group name (its subgroup label), indicator, student group size, score, year, Title I status, and grade span.

For a variety of reasons, including privacy and changes in grade spans, not all schools reported the same data in every year. Because a certain amount of data is required for regression models to be reliable, records for schools or student groups with insufficient data were dropped from the analysis. These exclusion criteria are listed in table B2. Further exclusions were made in the analysis of CSI and ATSI designations according to the procedure described in NJDOE's *2021-2022 Technical guide to Every Student Succeeds Act (ESSA) summative ratings and the identification of schools in need of support and improvement* (New Jersey Department of Education, 2023) for cases in which an insufficient amount of data is available to include a school in CSI or ATSI designations.

Table B2. Data exclusion criteria

Exclusion rule	Reasoning	Results
Indicator/group combinations for which six or fewer schools provided data	Data are insufficient for reliable model convergence.	All data on graduation rates for Native American students and multiracial students excluded, as well as chronic absenteeism data for Native American students
Schools reporting fewer than three years of data for a subgroup-indicator combination	Because the model uses a school's own data across multiple years, three years or more of data are required for reliable convergence.	About 4 percent of schools excluded from stabilization
Schools that changed grade spans by more than two grades and did not revert to the original grade span during the modeling period	Grade span changes are likely to reflect a significant change in how a school operates over the course of the modeling period. Such changes introduce bias (a meaningful, not random, difference) into the data between years, so using multiple years of the school's own data for stabilization purposes would be an unwise modeling choice.	About 5 percent of schools excluded from stabilization
Schools that reported data that could not be matched to a grade span	Because the accountability dataset did not include grade span information, that information came from other publicly available records provided by the New Jersey Department of Education (NJDOE). A small number of schools that reported accountability data could not be matched to grade span data.	<1 percent of schools excluded from stabilization
School/year/indicator/subgroup combinations for which only one observation of a specific school type (elementary, high school, or K-12) was available	NJDOE's Comprehensive Support and Improvement and Additional Targeted Support and Improvement designations rely on *z*-scores of composite indicators, weighted differently depending on school type; z-scores cannot be computed for a set of one.	<1 percent of data observations and no schools excluded from stabilization

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

The model

The study team fit a single stabilization model to each subgroup-indicator combination and then applied NJDOE's CSI and ATSI designation procedures to both the stabilized and unstabilized scores. The team fit each model separately to data from each subgroup. This approach helped avoid artificially inflating the precision of the estimates by treating subgroups as independent when they are likely to contain overlapping students (e.g., White students with disabilities).

The model is as follows:

$$y_{j,t} = \alpha + \alpha_j + \left(\beta + \beta_j\right)t + (\gamma + \gamma_j)C_t + \epsilon_{j,t}$$

$$\epsilon_{j,t} \sim N\left(0, \frac{\sigma^2}{n_{j,t}}\right)$$

where the estimated parameters are:

α is an overall intercept, representing the average intercept for all schools.

α_j is a school-specific intercept, representing the difference between overall performance for school j and the overall performance of schools on average.

β is an overall slope, representing the average change over time for all schools.

β_j is a school-specific slope, representing the difference between average change over time for all schools and change over time for school j.

C_t is an indicator variable that is 0 for years preceding the Covid-19 pandemic (years before 2020) and 1 for years during and after the pandemic.

γ is the overall effect of the Covid-19 pandemic, representing the average effect for all schools.

γ_j is the school-specific effect for the Covid-19 pandemic, representing the difference between overall Covid-19 effects and Covid-19 effects for school j.

The model includes the following priors:

$\alpha \sim N(0,1)$.

$\alpha_j \sim N(0, \sigma^2{}_{\alpha})$.

$\beta \sim N(0,1)$.

$\beta_j \sim N\left(0, \sigma^2{}_{\beta}\right)$.

$\gamma \sim N(0,1)$.

$\gamma_j \sim N\left(0, \sigma^2{}_{\gamma}\right)$.

$\sigma, \sigma_{\alpha}, \sigma_{\beta}, \sigma_{\gamma} \sim N^{+}(0,1)$.

The model includes $y_{j,t}$, the standardized unstabilized scores; $n_{j,t}$, the normalized student group size for each subgroup-indicator and year combination; and t, the centered year associated with the data.[4] Years were centered by labeling each year from 1 to 5 and subtracting the year label from the mean year label. Student group sizes were normalized to have a mean of 1. Standardizing the scores is a customary modeling decision that helps satisfy the assumptions of a standard normal likelihood, such as the presence of both negative and positive outcome values, a mean value of 0, and a standard deviation of 1 (see box B1 for assumptions of the model). Standardizing the scores also makes it possible to use recommended default prior distributions.[5] The model takes advantage of information about unstabilized scores for each subgroup-indicator combination in previous years and from other schools to estimate the trajectory of that subgroup-indicator combination in future years.

Box B1. Assumptions for this stabilization model

There are many ways to build a stabilization model. Choosing an effective model depends on the available data, computing resources, goals for the model, and reasonable assumptions about the data based on data exploration and subject matter expertise. For this study, the assumptions were as follows:

- For each indicator, a student group's performance in one year is likely to be predicted by its performance in other years and by statewide performance in that and other years. This is due to similarities in how indicators are defined and measured as well as shared experiences across groups of students with common characteristics.
- Changes in student group performance over time are likely to be small and therefore can be described well by a model that is linear in time.

[4] Throughout the report, *group* refers to a group of students within a school who fall within a specific subgroup. In our notation, the subscript j refers to school-specific terms.

[5] Not all indicators are well described by normal distributions. Although the standard priors selected for this model are designed to be gentle enough not to overwhelm the data, the model did not behave ideally for indicators that saturate near the bounds of the distribution, such as chronic absenteeism and graduation rates.

- Scores from the single year of data following the onset of the Covid-19 pandemic are likely to differ in a systematic way from scores from years preceding the onset of the pandemic.
- Scores are estimates of performance that are subject to error, that:
  - Has a larger effect on scores from smaller student groups.
  - Is random and unbiased and should therefore be normally distributed.

The study team fit models using Hamiltonian Monte Carlo as implemented in the Stan probabilistic programming language (Stan Development Team, 2021) via its R interface, RStan. Convergence and mixing were assessed using the Gelman-Rubin diagnostic and effective sample sizes (Gelman et al., 2013); nearly all Gelman-Rubin split $\hat{R}$ for parameters used to compute fitted values were less than or equal to 1.01.[6, 7] Similarly, across fits, all model parameters used to calculate fitted values had effective sample sizes of 100 or more.

After fitting the model, the study team used the mean of the posterior distribution, also called the stabilized scores, to answer research questions. In the interest of interpretability and clear parallels between stabilized and unstabilized scores, standard deviations and 95 percent credible intervals were not incorporated into the analysis. Because the model fits stabilized scores to a linear time trend and assumes normally distributed noise, a few stabilized scores fell slightly outside the standard 0 to 100 range. These scores were adjusted post hoc to 0 or 100.

Research question 1: Does stabilization behave as expected for each performance indicator?

To address research question 1, the study team compared stabilized and unstabilized scores and *z*-scores for all subgroup-indicator combinations in all years for which data were available. The team also compared the extent to which stabilization affected student groups of different sizes and different indicators. Simple criteria were established that can help decisionmakers identify indicators that are well-suited to stabilization.

Research question 2: Does stabilization improve the reliability of performance indicators, especially in groups of 10 to 19 students?

To address research question 2, the study team examined the extent to which stabilization reduced score variance across multiple student group sizes and how it affected the relationship between student group size and score variance. The study team also compared the representation of groups of 10 to 19 students and groups of 20 to 29 students in the extremes of the score distributions (below the 5th percentile or above the 95th percentile of score distributions) with their representation within the entire score distribution for both stabilized and unstabilized datasets. The study sought to determine whether groups of 10 to 19 students were more overrepresented in the extremes of unstabilized score distributions than groups of 20 to 29 students and whether stabilization improved representation for each.

[6] The recommended threshold for split $\hat{R}$ is 1.01; however, in some applications with larger numbers of estimated parameters, practitioners find it necessary to relax this threshold to 1.05. In the 66 models and thousands of parameters in the models for this study, only three parameters ($1.63x10^{-5}$ percent of model parameters) exceeded 1.01, and all were less than 1.03. Because it is not recommended to use split $\hat{R}$ alone, fits were assessed on both this and effective sample size parameters (Vehtari et al., 2021). Due to the number of models and their parameters, it was not feasible to present trace plots in this report, but the study team did visually inspect trace plots.

[7] Indicators for Native American students failed to converge due to extremely small group and sample sizes and were excluded from this assessment. The study team explored increasing iterations for this subgroup but found that this was not sufficient for convergence.

Research question 3: How does incorporating stabilized performance indicators into accountability designations change the set of schools designated as eligible using a designation process implemented without stabilization?

To address research question 3, the study team implemented NJDOE's CSI and ATSI designation process using both unstabilized data and data in which test-based indicators had been stabilized and then identified differences between the schools and student groups designated using each dataset. The stabilized dataset employed stabilization for test-based indicators but not for high school graduation rates and chronic absenteeism. This allowed the team to examine the effect that stabilization could have on CSI and ATSI designations if applied as recommended in this report. These analyses were conducted with 2018/19 data.

Limitations

The modeling choices used in this study were guided by the need to fit a model that would be accessible with readily available software and identifiable given the available data. The model selected for this study reflects the simpler of two approaches that the study team explored to capture the effects of the Covid-19 pandemic on scores. In many cases, scores in pandemic-affected years were outliers, so fitting a simple linear model risked the pandemic-affected scores overwhelming the pre-pandemic scores and having an outsized impact on stabilized scores.

To address this risk, the study team first explored fitting a model that allowed for nonlinear trends in scores but found that this model could not be consistently estimated without more data than were available. The model selected for this study fits a linear time trend to pre-pandemic scores and a separate Covid-19 effect to scores from the year following pandemic onset. Although this model consistently met the study team's standards for convergence, the results are influenced somewhat by the fact that only one year of data affected by the Covid-19 pandemic was available.

This had several effects. First, the pandemic disrupted accountability processes, so data for indicators for math and English language arts growth were unavailable for the 2021/22 school year. Second, because of how the model accounts for Covid-19's effect on scores, the linear component of the model relied on a maximum of four years of data rather than five for each subgroup-indicator combination. This can result in less reliable and more exaggerated estimates of linear time trends. Additionally, the separate Covid-19 effect was challenging for the model to estimate using only one year of data.

As more data become available for years after the onset of the pandemic, states may find that they have different concerns about how to best account for the effect of Covid-19 on their own data.

Additionally, because the data available were at the subgroup level rather than the student level, subgroup-indicator combinations are stabilized independently. A group's stabilized scores are influenced by its own performance in other years and the performance of other student groups in that subgroup-indicator combination, but not by its own performance on other indicators or the performance of other groups in the same school. A stabilization model built with student-level data would be able to draw on a much richer set of information and likely provide more accurate performance estimates.

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

New Jersey Department of Education. (2023). *2021-2022 Technical guide to Every Student Succeeds Act (ESSA) summative ratings and the identification of schools in need of support and improvement.* https://www.nj.gov/education/title1/accountability/docs/22/2021-22_ESSA_Technical_Guide_SummativeRatings_Identification.pdf

New Jersey Department of Education. (2024). Accountability. Title I, Part A. https://www.nj.gov/education/title1/accountability/

Stan Development Team. (2021). *Stan modeling language user's guide and reference manual, 2.30*. https://mc-stan.org

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis,* 16(2), 667-718. https://doi.org/10.48550/arXiv.1903.08008

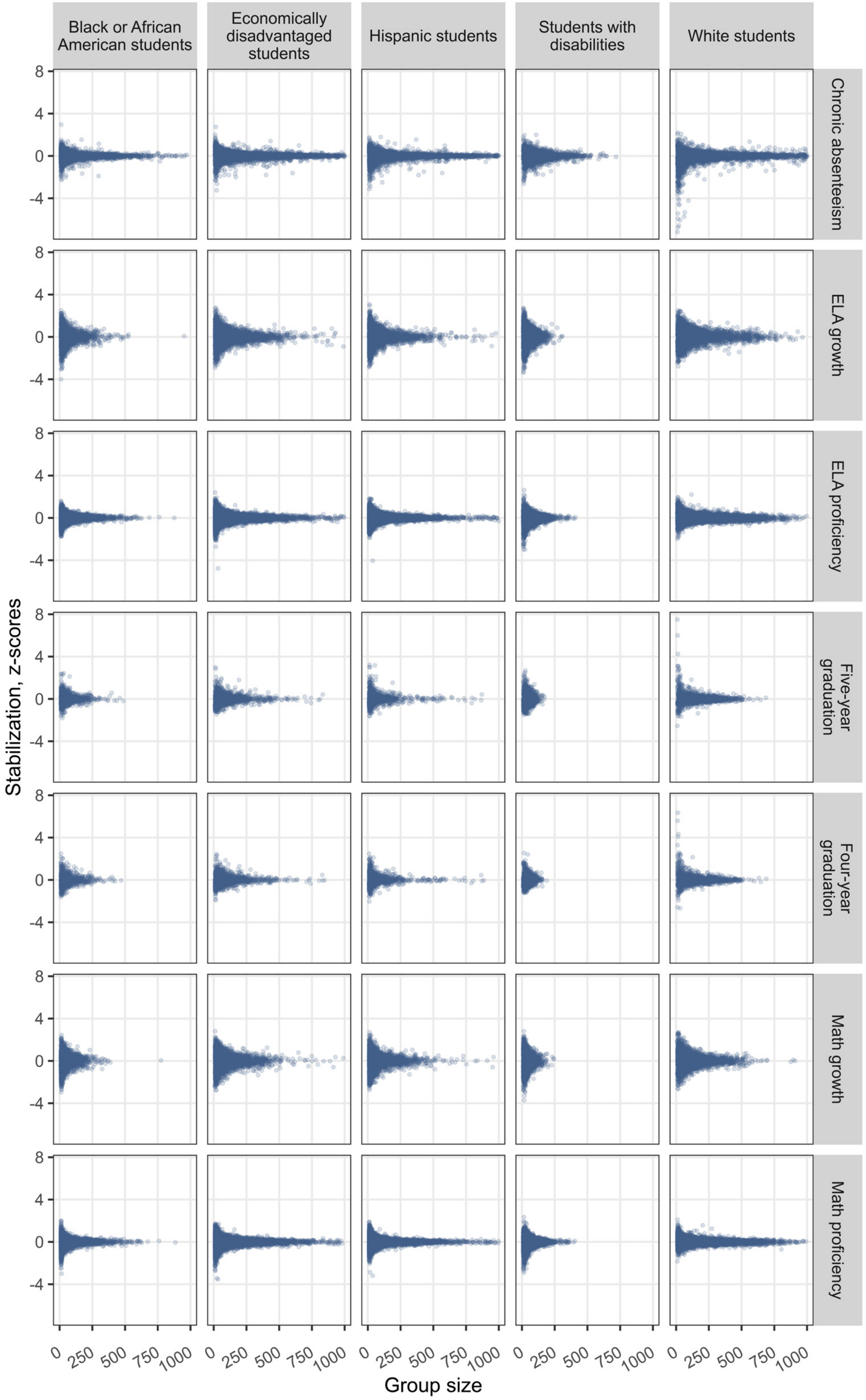
Appendix C. Supporting analysis

This appendix presents supporting information for the findings in the main report, including visualizations equivalent to those presented for findings in the main report for the five largest subgroups in New Jersey and for all indicators except English language proficiency, which does not apply to those five subgroups. The study team chose to display findings only for the five largest subgroups to improve the readability of the images, but result patterns were consistent across all subgroups. Note that throughout this appendix, some axis units will differ across subgroups and indicators to improve the readability of these small images.

Stabilization has a larger effect on smaller groups and extreme scores

Stabilization has a larger effect, on average, on smaller student groups (figure C1), and small stabilization effects are, in general, more common than large ones (figure C2). This is consistent across all subgroup-indicator combinations for which the distribution peaks at zero. A logarithmic scale on the y axis improves visibility for small counts.

Figure C1. Stabilization of *z*-scores by student group size for the five largest subgroups

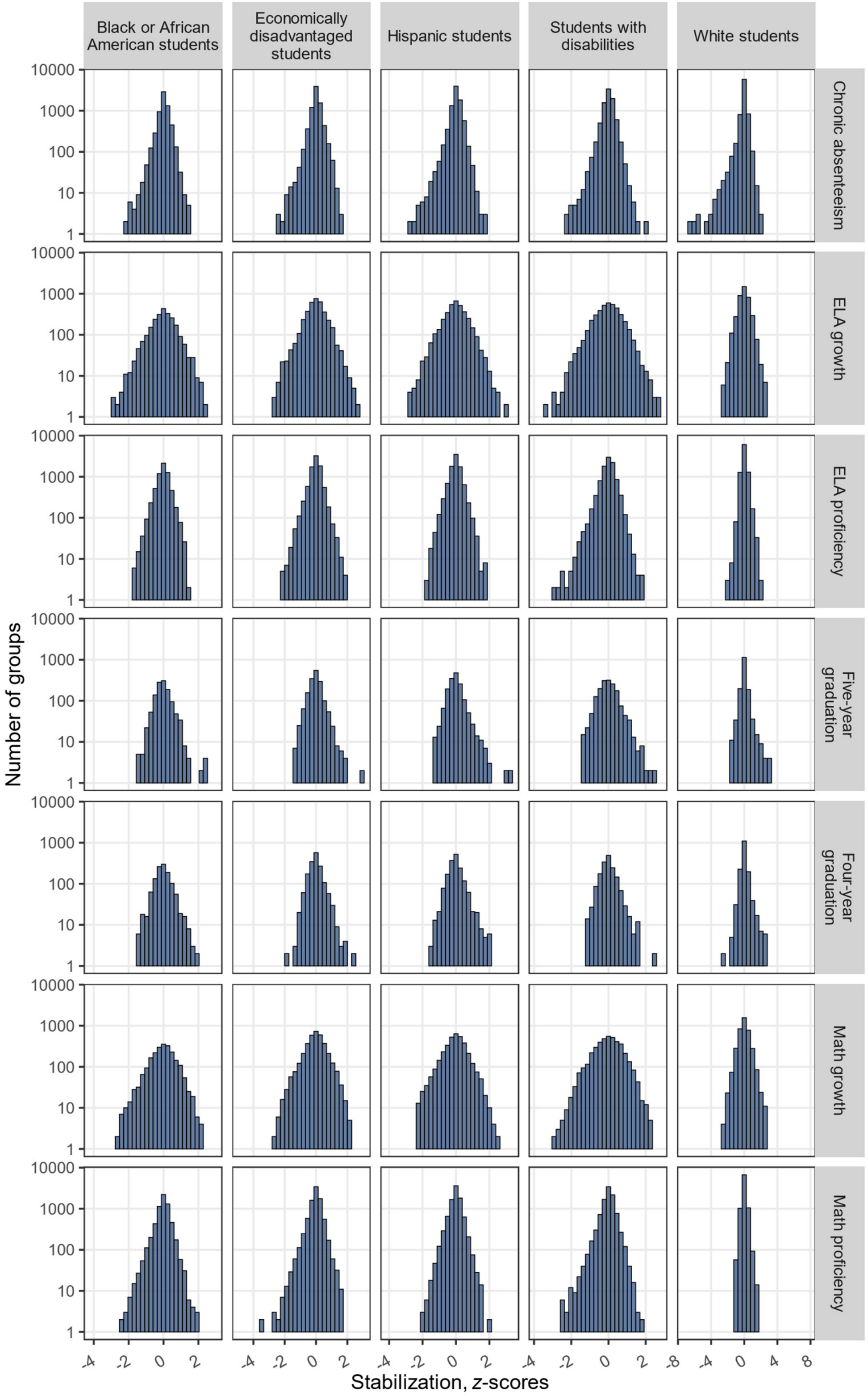


ELA is English language arts. ELP is English language proficiency.

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for proficiency indicators, between 2017/18 and 2021/22 for ELP progress, between 2016/17 and 2018/19 for growth indicators, between 2015/16 and 2021/22 for high school graduation rates, and between 2016/17 and 2021/22 for chronic absenteeism. The analysis covers the five largest subgroups and all indicators except English language proficiency, which does not apply to those five subgroups.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Figure C2. Distribution of stabilization effects by number of groups for the five largest subgroups

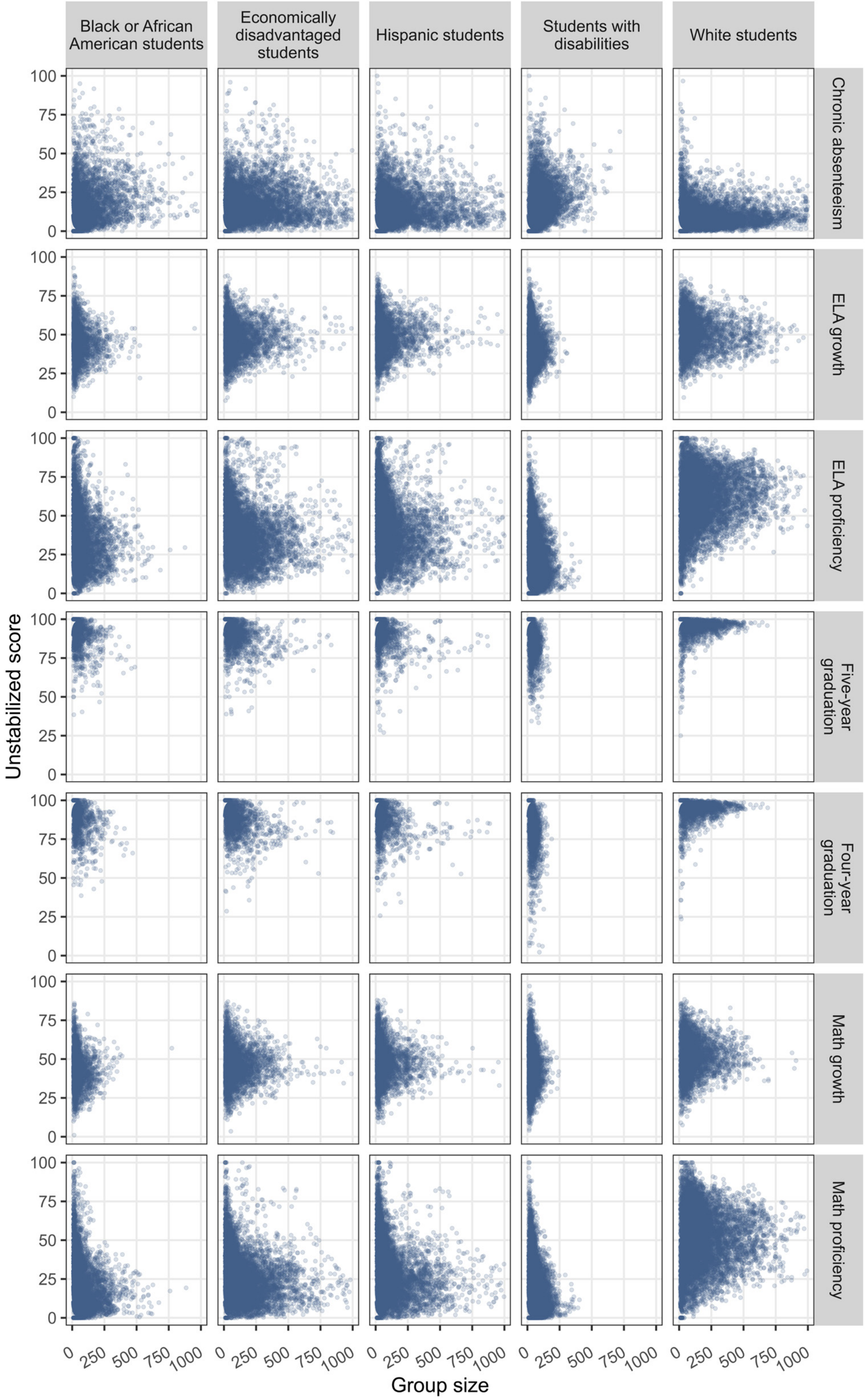


ELA is English language arts. ELP is English language proficiency.

Note: Stabilized and unstabilized score data. The analysis covers the five largest subgroups and all indicators except English language proficiency, which does not apply to those five subgroups. Data were collected between the school years 2015/16 and 2021/22 for proficiency indicators, between 2017/18 and 2021/22 for ELP progress, between 2016/17 and 2018/19 for growth indicators, between 2015/16 and 2021/22 for high school graduation rates, and between 2016/17 and 2021/22 for chronic absenteeism.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Figure C3. Unstabilized scores by student group size for the five largest subgroups



ELA is English language arts. ELP is English language proficiency.

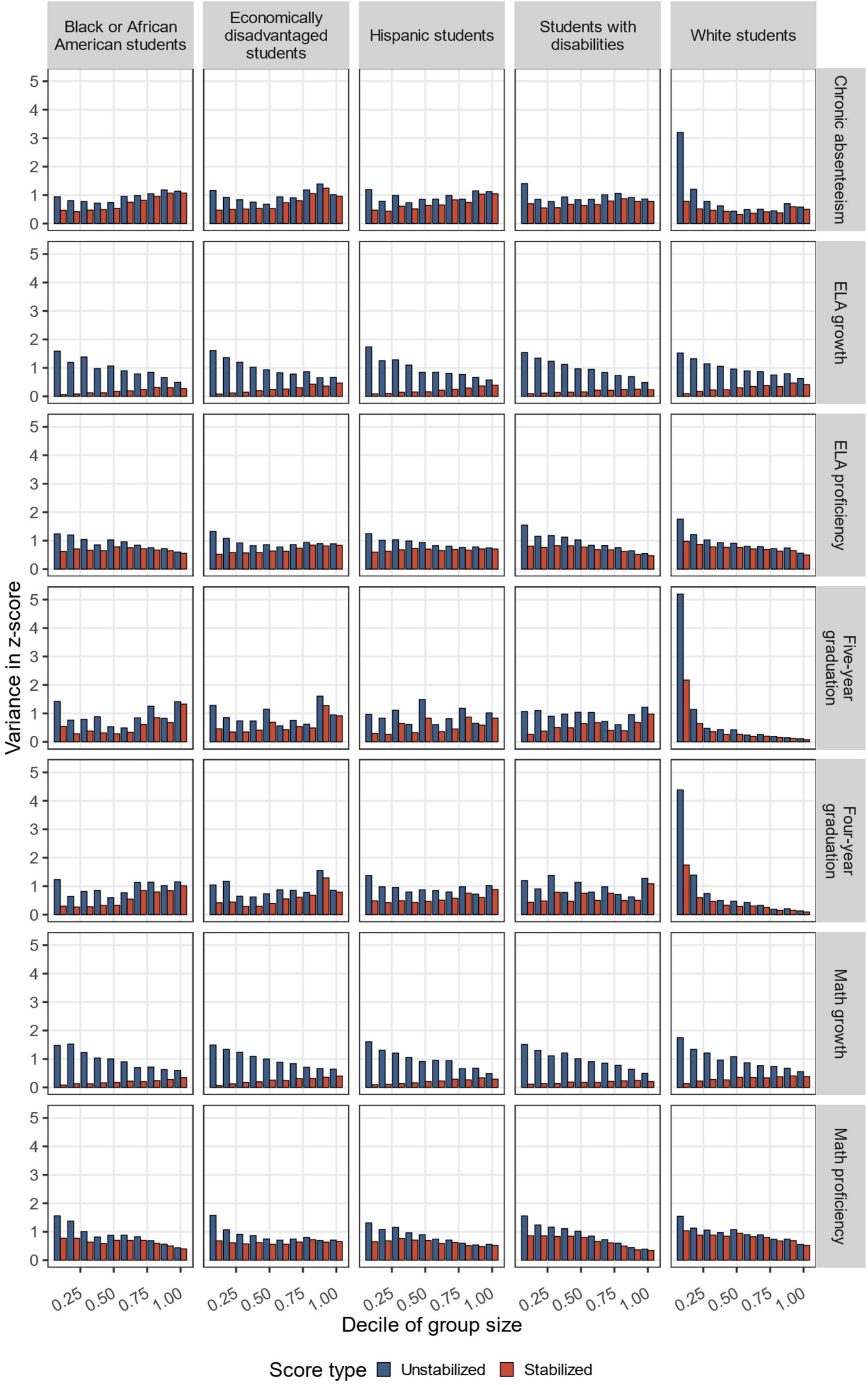
Note: Stabilized and unstabilized score data. The analysis covers the five largest subgroups and all indicators except English language proficiency, which does not apply to those five subgroups. Data were collected between the school years 2015/16 and 2021/22 for proficiency indicators, between 2017/18 and 2021/22 for ELP progress, between 2016/17 and 2018/19 for growth indicators, between 2015/16 and 2021/22 for graduation rates, and between 2016/17 and 2021/22 for chronic absenteeism.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Stabilization can reduce the relationship between student group size and indicator variance

Across all subgroups and indicators, stabilization nearly always reduces the variance of scores within group size deciles, and differences in variance between stabilized and unstabilized distributions are larger for smaller student groups, as expected (figure C4). For test-based indicators, variance is generally more uniform after stabilization, which is indicative of a reduced relationship between student group size and score variance. This is less consistent for non-test-based indicators, where the relationship between student group size and score variance may be more uniform and noninverse.

Figure C4. Variance of *z*-scores by student group size deciles for the five largest subgroups



ELA is English language arts. ELP is English language proficiency.

Note: Stabilized and unstabilized score data. The analysis covers the five largest subgroups and all indicators except English language proficiency, which does not apply to those five subgroups. Data were collected between the school years 2015/16 and 2021/22 for proficiency indicators, between 2017/18 and 2021/22 for ELP progress, between 2016/17 and 2018/19 for growth indicators, between 2015/16 and 2021/22 for graduation rates, and between 2016/17 and 2021/22 for chronic absenteeism.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Appendix D. Other analyses: Implications of the use of group-level data arising from relationships among demographic composition, subgroup size, and scores

Stabilization models are, by design, skeptical of extreme scores from small student groups and will tend to adjust extreme scores toward the mean score, with larger adjustments for smaller student groups. Because the data available for this study were at the subgroup level and could not account for overlap among student groups, the model chosen for this study stabilizes scores from student groups by borrowing information from other student groups within a subgroup-indicator combination and from the same student group across time. Therefore, the model can only partially account for differences in demographic composition between schools. That is, models for subgroup-indicator combinations are necessarily independent of one another, so the model cannot borrow information from different student groups in the same school, which is likely to be very predictive of performance. This is because students within the same school share teachers, administrators, school resources and, often, other important factors such as socioeconomic status.[8]

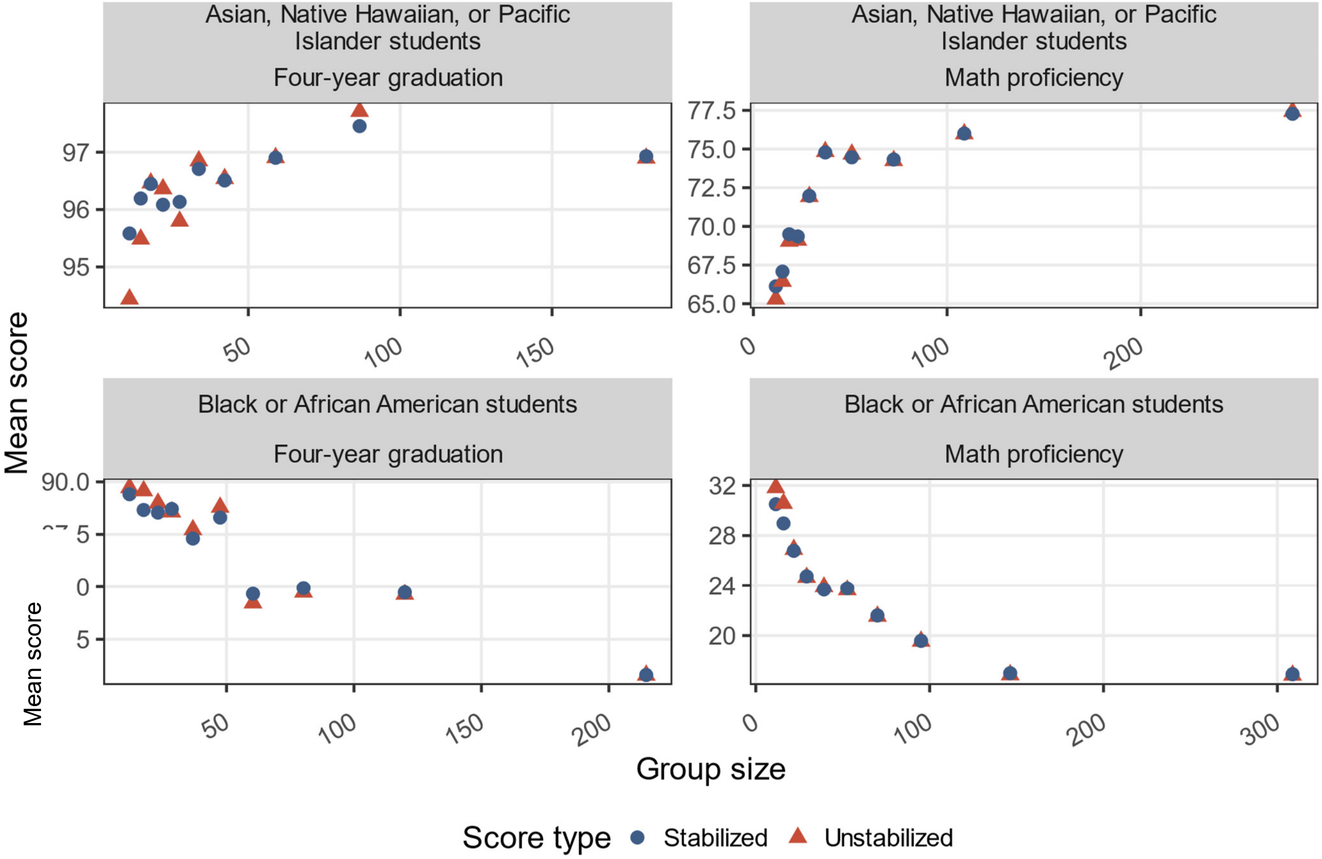
The use of subgroup-level data results in the implicit limiting assumption that a student group's performance on an indicator is predicted by other student groups within the same subgroup but not by other student groups within the same school. For example, because a stabilization model will often assume that there are not meaningful differences in the true performance of students in groups of different sizes, the scores from small student groups will tend to be stabilized toward the overall mean, with some bias toward the mean for larger student groups. If scores tend to improve or decline as a function of student group size (which may reflect known predictive factors of school performance, such as location), this could undermine the quality of the stabilized results.

The model lessens the effect of this limitation by assigning group-specific effects for each student group modeled, so that the stabilized results for a student group are informed by that group's previous performance and, therefore, by the characteristics of the school to which the student group belongs.

To explore the extent to which this poses a practical challenge for the work conducted for this study, the mean stabilized and unstabilized scores on each indicator were calculated for each subgroup by decile of group size. The results (plotted for focal subgroup-indicator combinations in figure D1) show that, for several subgroup-indicator combinations, there are clear relationships between student group size and score. However, these relationships are well preserved by the stabilized results, indicating that stabilization introduces little bias in results and may be used with minimal concern in this regard.

[8] Demographics are also a potential source of measurement error. Hermann et al. (2016) found that test scores for students who are economically more disadvantaged also have more measurement errors than test scores for students who are economically less disadvantaged.

Figure D1. Relationships between student group size and unstabilized scores are well preserved in the stabilized results, indicating that stabilization introduces little bias in related results



Note: Stabilized and unstabilized score data. Data were collected between the school years 2015/16 and 2021/22 for proficiency indicators and between 2015/16 and 2021/22 for graduation rates. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Reference

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy, 3*(1), 1-10. https://doi.org/10.1080/2330443X.2016.1182878